Nucleic Acids Research, 2005, Vol. 33, No. 3 ***e31***
doi:10.1093/nar/gni027

A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array

Jeremy Harbig, Robert Sprinkle and Steven A. Enkemann*

H. Lee Moffitt Cancer Center and Research Institute and the University of South Florida Tampa, Florida 33612, USA

Received November 3, 2004; Revised January 10, 2005; Accepted January 20, 2005

ABSTRACT

One of the biggest problems facing microarray experiments is the difficulty of translating results into other microarray formats or comparing microarray results to other biochemical methods. We believe that this is largely the result of poor gene identification. We re-identified the probesets on the Affymetrix U133 plus 2.0 GeneChip array. This identification was based on the sequence of the probes and the sequence of the human genome. Using the BLAST program, we matched probes with documented and postulated human transcripts. This resulted in the redefinition of approximately 37% of the probes on the U133 plus 2.0 array. This updated identification specifically points out where the identification is complicated by cross-hybridization from splice variants or closely related genes. More than 5000 probesets detect multiple transcripts and therefore the exact protein affected cannot be readily concluded from the performance of one probeset alone. This makes naming difficult and impacts any downstream analysis such as associating gene ontologies, mapping affected pathways or simply validating expression changes. We have now automated the sequence-based identification and can more appropriately annotate any array where the sequence on each spot is known.

INTRODUCTION

The current dogma suggests that microarray data is erratic and poorly reproducible. It is recommended by the scientific community that all genes identified in a microarray experiment be verified by a more commonly accepted method such as RT–PCR or northerns (1). Yet there is significant evidence that microarray data is highly reproducible (2–7). Why is it that the technology is highly reproducible within one format, but less reproducible when working across methodologies? A closer look at reports, where microarray results were poorly reproduced or inconsistent with other methods, suggests that the fault does not lie with the biochemical methods, but rather with the bookkeeping (8,9). For example, experiments using spotted cDNAs are dependent on the accurate maintenance of the bacterial stocks that house the DNA eventually used on spots. This is not always done effectively. Some arrays can have as many as 30% of their spots misidentified because of errors in the DNA stocks (9–11). Because of this, many spotted arrays now use sequence-verified clones or synthesized oligos (12–14). However, this does not remove all possible sources of misidentification.

The alternative to spotted microarrays has been the *in situ* synthesized oligonucleotide arrays marketed by the Affymetrix Corporation (15). This format has less chance for error since the sequence produced on each spot is known. Yet, even this format can be plagued by incorrectly identified spots (8,16). Part of the problem is that the probes on an array are identified based on what the company was hoping to detect, not based on what they actually detect. There are two reasons why these are not the same thing. One is the concept that each spot should detect a single gene; the second is that there are often problems with the sequences upon which the probes are based. One example of this latter problem is illustrated by the probeset 214019_at, found on the Human Genome U133A chip and the U133 plus 2.0 arrays. The probes in this probeset were designed based on the GenBank sequence Z23022. According to the description of this gene at the NetAffx annotation support site for Affymetrix, this probeset identifies the transcript for cyclin D1. But it does not. The sequence Z23022 is a hybrid sequence and does not represent an actual cellular transcript. Part of this GenBank sequence comes from the cyclin D1 gene, which is located on chromosome 11; the rest of the sequence is derived from the tip of chromosome 19. The probes synthesized on each array are designed from the chromosome 19 portion of this hybrid sequence while the definition of the gene comes from the chromosome 11 portion of the sequence. Therefore, the annotation describes this probeset as cyclin D1 although the probes instead detect a mildly repetitive sequence in the

*To whom correspondence should be addressed. Tel: +1 813 745 9033; Fax: +1 813 979 7265; Email: enkemasa@moffitt.usf.edu human genome that has retroviral characteristics. This ERVK element is repeated hundreds of times in the human genome with the closest copy over 1.9 million bases downstream of the cyclin D1 gene.

The fact that a probe sequence can detect more than one gene is a more pervasive problem. A hybrid clone, like the one described above, would certainly detect more than one gene if it were used as a cDNA spot. However, even short oligonucleotides can detect more than one gene. Many primary transcripts are alternatively spliced under different growth conditions or in different cell types. According to the original definition of a gene, these alternative transcripts should be considered different genes. They are more commonly referred to as splice variants. Either way, it is uninformative to consider all splice variants as equivalent. Consider the case of the FLICE-inhibitory proteins. These are encoded from the same locus, but some isoforms are pro-apoptotic and others are anti-apoptotic (17). It is essential that one knows which isoforms are expressed in order to fully understand the implications of an increase or decrease observed in a microarray experiment. An oligonucleotide, or other probe, that detects several splice variants is not as informative as one that detects a single species of transcript. A proper probe identification should indicate which case applies. A second reason that a probe may detect more than one gene is the occurrence of duplicated genes and gene families. There are numerous instances of closely related genes, and many times cross-hybridization to DNA probes has been used to identify these evolutionarily related sequences. In microarrays, this cross-hybridization can be a problem unless it is properly identified.

We have utilized the current knowledge of the human genome to make a more correct identification of the probesets on an Affymetrix array. Using the sequence of the probes as the starting point, we have re-identified the genes detected by the commonly used U133 plus 2.0 array. Our final annotation indicates where the identification is questionable or complicated by cross-hybridization issues. This more realistic annotation of spots on a microarray will allow individuals to more accurately translate finding to other microarray formats and design informative follow-up experiments.

METHODS

Our knowledge of the human genome is sufficiently mature to allow it to form the reference state for the identification of genes. The RefSeq initiative at the National Center for Biotechnology Information (NCBI) has been generating model RNA molecules that represent individual transcripts of human genes (18,19). We made use of these sequences, and the locus from which they are derived, to identify genes that are interrogated by individual probes on Affymetrix GeneChips. Each probeset on a GeneChip consists of 22–48 oligonucleotide probes 25 bases long. Half of these are designed to be a perfect match for a specific transcript. The other half are designed with a single base mismatch in the center of the 25 base oligonucleotide. Each cumulative set of probes is designed to detect a selected region of a possible transcript. This sequence is referred to as the ‘target’ for the probes. These targets are available from Affymetrix through their NetAffx web site (20). We used these targets for the initial survey of possible transcripts that each probeset might detect. We used the BLAST program to search GenBank for any human RefSeq that matched the target (21). In the absence of a human RefSeq, we accepted any human sequence that matched the target. This initial search used the blastn program against the nr database using a word-size of 28. The returned results were collected and screened for sequences pertaining to cDNAs or mRNAs and gave a score greater than 100. These results were stored in a local database associated with the probeset ID number from the chip. A second screen was performed where each individual probe within a probeset was compared with the sequences collected from the first screen. Probes were scored based on how many bases exactly matched the retrieved sequences. The sequence with the best average score across all probes in a probeset was chosen as the best match. This serves as the primary identification for the probeset. However, in many cases the probes equally detected several reference sequences. When this occurred, a probeset was listed as detecting multiple transcripts. We have also indicated transcripts that scored lower, but might also hybridize to probes in the probeset and therefore contribute to the signal measured.

The probesets were also evaluated for a number of potential complicating factors. Each probe on the array was compared to the 8 ALU reference sequences that represent the most common members of this repetitive element commonly found in human transcripts: Genbank accession numbers U14567, U14568, U14569, U14570, U14571, U14572, U14573 and U14574 (22). Probes that can hybridize to these sequences are flagged. The database was also used to identify split or chimeric probesets. If some of the probes in a probeset detected one sequence while the remainder detected a distinctly different sequence, the probeset was selected for further review. These probesets were then manually evaluated to distinguish between probesets that detected two distinct gene products and those detecting possible splice variants, recognizing complex loci such as the immunoglobulin genes, or recognizing repetitive elements. Any probeset that failed to find a matching sequence through any of the above techniques was manually identified by comparison to all available GenBank sequences until a match was found.

The database of annotation is available at http://mriweb.moffitt.usf.edu/mpv/. Details of the computational methods and the database are also available online.

RESULTS

The Affymetrix gene chip U133 plus 2.0 contains 54 675 probesets. There are 62 probesets designed for special functions, such as measuring supplementally added transcripts. These probesets have the prefix AFFX. This leaves 54 613 probesets designed specifically for the detection of human genes. We re-evaluated the identification of these probesets using the sequence of each probe as the basis for the identification. This reassessment resulted in the renaming of 20 415 probesets. The revised identifications are available through the link http://mriweb.moffitt.usf.edu/mpv/. Based on our new identification of the probes on this array, we estimate that the U133 plus 2.0 gene chip can detect more than 30 000 human transcripts derived from more than 20 000 loci within the human genome. This does not mean that in any single experiment it would be possible to identify 30 000 genes. Many probesets detect multiple genes and thus one may not be able to discern which of several possible transcripts gave rise to the hybridization signal.

Cross-hybridization affects gene identification

Cross-hybridization is an issue that was considered in the design of the GeneChip arrays. The mismatch probes are intended to capture non-specific cross-hybridization, so that it can be accounted for in the analysis of the data. However, we found 206 mismatch probes that are a perfect match for some human transcript. In most cases, there is only one mismatch probe affected per probeset and thus only a few calculations will be adversely affected by this phenomenon. A bigger problem is specific cross-hybridization, where the perfect match probes can recognize more than one transcript. We found over 5000 probesets that specifically hybridize to more than one cataloged splice-variants or transcripts from another locus. We have selected a few examples to illustrate the complexity of this problem.

Gene families make up a significant portion of the human genome. On the p arm of chromosome 9 there are 13 closely related genes for interferon alphas (23). There are 13 probesets on the U133 plus 2.0 chip that detect these genes. Table 1 shows that the closest match to these probesets, based on the sequence of the probes, is not always the same species as indicated by the Affymetrix annotation. This illustrates that the best intentions of probe design are not always realized when the experiment is performed. Our evaluation indicates that there is considerable cross-hybridization expected between these related transcripts and that the best match family member is not the only transcript that can hybridize to these probes. This point is best illustrated by the fact that the exact same probe sequence (AGAAATACAGCCCTTGTGCCTGGGA) is used as one of the probes for seven different probesets from this group. Two other probesets contain a closely related overlapping sequence. There is no question that cross-hybridization will occur when the sequence is exact, but weaker matches must also be considered. When viewing a spot on a microarray, one cannot tell the difference between high expression combined with weak cross-hybridization and low expression but perfect hybridization. Therefore, it would be difficult to identify exactly which gene is responsible for the signal detected without considering all 13 probesets as a group. Perhaps viewing individual probes, reassigning probes to novel probesets, or comparing the relative expression levels between probesets can help one determine exactly which gene is expressed. At a minimum, it is important to understand that something more is required to interpret information related to these genes and to properly design follow-up experiments. One cannot evaluate these probesets as individuals and get conclusive results.

A second example relates to the protocadherin gamma family of genes. This family of genes has multiple variable first exons, but share their 3′ most exon (24,25). Since most probesets are designed to detect the 3′ regions of transcripts, they do not distinguish between individual family members. The probesets 205717_x_at, 209079_x_at, 211066_x_at and 215836_s_at detect all 22 different protocadherin gamma family members because they recognize this shared terminal exon. A few other probesets recognize some of the unique first exons, but there are not enough probes to identify all members of this gene family. One is therefore uncertain which

Table 1. Probesets that detect members of the interferon alpha gene family

Probeset ID	Probe sequence	Probe location		Probeset member	Best match (score)	Sequence-based ID	Affymetrix reference sequence	Affymetrix ID
		X	Y					
211405_x_at	AGAAATACAGCCCTTGTGCCTGGGA	514	123	Probe 10	NM_021268 (24.1)	Interferon, alpha 17	NM_002170	Interferon, alpha 8
207964_x_at	AGAAATACAGCCCTTGTGCCTGGGA	515	123	Probe 6	NM_021068 (25.0)	Interferon, alpha 4	NM_021068	Interferon, alpha 4
208182_x_at	AGAAATACAGCCCTTGTGCCTGGGA	516	123	Probe 3	NM_002172 (25.0)	Interferon, alpha 14	NM_002171	Interferon, alpha 10
208259_x_at	AGAAATACAGCCCTTGTGCCTGGGA	517	123	Probe 8	NM_021057 (25.0)	Interferon, alpha 7	NM_002175	Interferon, alpha 21
211145_x_at	AGAAATACAGCCCTTGTGCCTGGGA	518	123	Probe 9	NM_002175 (23.6)	Interferon, alpha 21	NM_002175	Interferon, alpha 21
208344_x_at	AGAAATACAGCCCTTGTGCCTGGGA	519	123	Probe 10	NM_006900 (25.0)	Interferon, alpha 13	NM_024013	Interferon, alpha 1
208448_x_at	AGAAATACAGCCCTTGTGCCTGGGA	520	123	Probe 3	NM_002173 (25.0)	Interferon, alpha 16	NM_002171	Interferon, alpha 10
208261_x_at	AGGAAATACAGCCCTTGTGCCTGGG	17	79	Probe 3	NM_002171 (25.0)	Interferon, alpha 10	NM_002171	Interferon, alpha 10
208548_at	AGAGAAAAAGTACAGCCCTTGTGCC	248	113	Probe 10	NM_021002 (25.0)	Interferon, alpha 6	NM_000605	Interferon, alpha 2
207932_at	No overlapping probe				NM_002170 (25.0)	Interferon, alpha 8	NM_002170	Interferon, alpha 8
208375_at	No overlapping probe				NM_024013 (25.0)	Interferon, alpha 1	NM_024013	Interferon, alpha 1
211338_at	No overlapping probe				NM_000605 (23.6)	Interferon, alpha 2	NM_000605	Interferon, alpha 2
214569_at	No overlapping probe				V00541 (25.0)	Interferon, alpha 5	NM_002169	Interferon, alpha 5

Indicated are single probes from several probesets that are highly similar or identical, their location on the array, and the probe number from an 11 probe set. Also indicated is the most similar gene to the probes in the probeset with the score indicating the average match across 25 possible nucleotides for 11 probes. The last two rows contain the Affymetrix reference sequence and the gene name indicated at their NetAffx website.

transcripts might be contributing to the behavior that led one to find these probes in their microarray experiment.

Myosin 18A and TGFβ1-induced apoptotic factor 1 (TIAF1) represent a similar situation. These two genes share a common final exon. Myosin 18A is a large transcript composed of more than 30 exons and there are two splice variants encoding slightly different proteins. Both splice variants share a final exon with the 3′ end of TIAF1. In this case, there is only one probeset on the Affymetrix U133 plus 2.0 chip (202039_at) and it recognizes this common 3′ region. Therefore, there is no mechanism for distinguishing which gene is hybridizing to an array. Affymetrix has designated this probeset as recognizing TIAF1. A more proper designation should indicate that it detects three possible transcripts so that the appropriate gene can be eventually determined.

Some probes do not detect defined human genes

There are a number of probesets that do not recognize anything in GenBank that might be a human mRNA and do not recognize anything in the current assembly of the human genome. These probesets are listed in Table 2. Many of these do not detect any transcript, but a few are convenient because of what they do detect. Probeset 221106_at recognizes a rat gene and does not cross-hybridize with anything in the human genome. Although there is a corresponding human gene, it is sufficiently different from the designed probes that the level of cross-hybridization is expected to be minimal. There is very little hybridization detected by this probeset in any tissue and certainly much less than from the better designed probeset, 218675_at (data not shown). This probeset can help form a basis for how much cross-hybridization one might expect from weakly similar transcripts. Of more utility with respect to cross-hybridization is the probeset 217680_x_at. Affymetrix lists this probeset as recognizing the ribosomal protein L10. As a probeset it does not, although some of the individual probes do. These probes were also designed based on a chimeric EST (expressed sequence tag) sequence and the probes span the junction between ribosomal protein L10 sequences and vector sequences. The probes that do recognize ribosomal protein L10 provide an interesting look at the cross-hybridization issues facing this technology. Table 3 shows that as individual probes match more of the transcript for ribosomal protein L10, the signal derived from hybridization increases. Probe 11 matches the L10 transcript across all 25 bases of the oligo. Since ribosomes are essential components of cells, the L10 transcript is constitutively and highly expressed (26). This probe has a high signal in all arrays hybridized with human samples. For comparison purposes, compare the signal derived from probeset 200725_x_at or the individual probes that comprise this probeset. In contrast, probes 10 through 7 match progressively less of the ribosomal protein transcript. There is also a corresponding decrease in the signal intensity recorded by these probes. This decreased signal continues down to the level expected from spurious cross-hybridization. Probes 1 and 2 from this probeset do not detect the ribosomal transcript, but possibly hybridize to a transcript from LOC388732. The values in Table 3 are the average of 25 samples selected at random. Two other probesets detect plasmid-derived sequences rather than human transcripts. Probeset 211371_at detects the bovine growth hormone polyA sequence incorporated into some expression vectors. The

Table 2. Probesets that do not detect human genes

Probeset ID	U133 plus 2.0 library description	NetAFFX identification	Blast identification	Manual identification
214089_at	Mitogen-activated protein kinase kinase kinase kinase 3	Ribosomal protein S8	No match found	Detects no gene
214379_at	Heterogeneous nuclear ribonucleoprotein D-like	Heterogeneous nuclear ribonucleoprotein D-like	No match found	Detects no gene
214689_at	Pregnancy-associated plasma protein-E	Placenta-specific 3	No match found	Detects no gene
214935_at	Hypothetical protein	Nucleoporin 62 kDa	No match found	Probes 9–11 weakly hybridize to nucleoporin 62 kDa transcripts
217680_x_at	EST	Transcribed sequence with strong similarity to 60S ribosomal protein L10	No match found	Some probes hybridize to ribosomal protein L10 and similar transcripts (best match LOC284393)
217712_at	Moderately similar to ALU8	Transcribed sequence with weak similarity to cytokine receptor-like factor 2	No match found	The gene is not yet defined. The probes recognize a sequence repeated 6 times on the X chromosome.
222181_at	CCR4-NOT transcription complex, subunit 2	CCR4-NOT transcription complex, subunit 2	No match found	Probe 5 recognizes CCR4-NOT transcription complex, subunit 2. Probes 2-4 also bind weakly
220932_at	Hypothetical protein	No ID	No match found	Detects no gene
211371_at	MAP kinase kinase MEK5c	Mitogen-activated protein kinase kinase 5	U71088	Bovine growth hormone poly A sequence engineered into commercial cloning vectors
222227_at	Zinc finger protein 236	Zinc finger protein 236	AK000847	SV40 poly A sequence engineered into commercial cloning vectors
221106_at	HBOIT for potent brain type organic iontransporter	Solute carrier family 22 (organic cation transporter), member 17	No match found	Rattus norvegicus solute carrier family 22
214019_at	BCL1 mRNA encoding cyclin	Cyclin D1 (PRAD1: parathyroid adenomatosis 1)	Z23022	Mildly repetitive endogenous retroviral like element (ERVK)

Shown is a comparison of the identifications from the original definition of the probeset, the current definition available at Affymetrix, our definition based on the sequence of the probes and a manual identification intended to define where the original sequence came from.

Table 3. Characteristics of probes within the probeset 217680_at

Probe number	Similarity to ribosomal protein L10	PM value	MM value
1	No match	452	131
2	No match	346	702
3	No match	31	62
4	No match	52	99
5	No match	21	24
6	No match	185	195
7	13 base match	375	326
8	16 base match	524	525
9	20 base match	6727	2134
10	23 base match	24 927	2607
11	25 base match	41 487	18 066

The last five probes detect the ribosomal protein L10 with increasing affinity. The probe values for the perfect match probes (PM) and the mismatch probes (MM) are the average of 25 independent chip measurements.

probeset 222227_at detects the more commonly used SV40 polyA tail region. This is a very useful probeset because it detects many of the expression vectors used for transient and stable transfections. It can detect samples derived from transfected cells and may be useful for identifying SV40 infected tissues. Changes in the expression level of this probeset might be useful in special circumstances, but for most experiments the probesets listed in Table 1 should be removed from further consideration before performing an analysis of the data.

We found 448 probesets that were the inverse complement of the intended target gene. These probesets appear to have been designed based on poorly defined sequences in which the orientation was not defined and the gene names may have been assigned later. As the inverse complement, the probes will not detect the indicated gene. Not all of these probesets are useless. One of the interesting observations about the human genome is that it is very common to find genes on opposite strands of the chromosome arranged in tandem. Therefore, a gene is arranged head-to-head with one adjacent neighbor and probably shares a common promoter region and is arranged tail-to-tail with the other neighbor. Genes arranged tail-to-tail often overlap their 3′ untranslated regions. Since many of the Affymetrix probes are designed to detect the 3′ untranslated regions, a probeset can detect one of these genes and be the inverse complement of the other. Therefore, many of the inverted probesets have simply been assigned a new name. In many cases, an adjacent gene could not be identified. Although a name cannot be assigned, we can attempt to determine if the probes are detecting a transcript. As a screen for presumptive transcripts, we simply looked at the expression level across more than 500 different microarrays generated from more than 130 different organs, tissues and cell lines. In the interest of clarity, we have only presented the results of 100 samples representing more than 50 different cell lines and tissues. Figure 1 shows the expression level measured across these 100 samples for two probesets that appear to detect transcripts and one probeset that does not appear to detect a transcript. Since we have not tested all possible cell types, it is not possible to conclude that probesets such as 216868_s_at do not detect anything. Nonetheless, they should be considered suspect in any experiment. The probeset 213619_at appears to detect a transcript in many different cells and tissues although

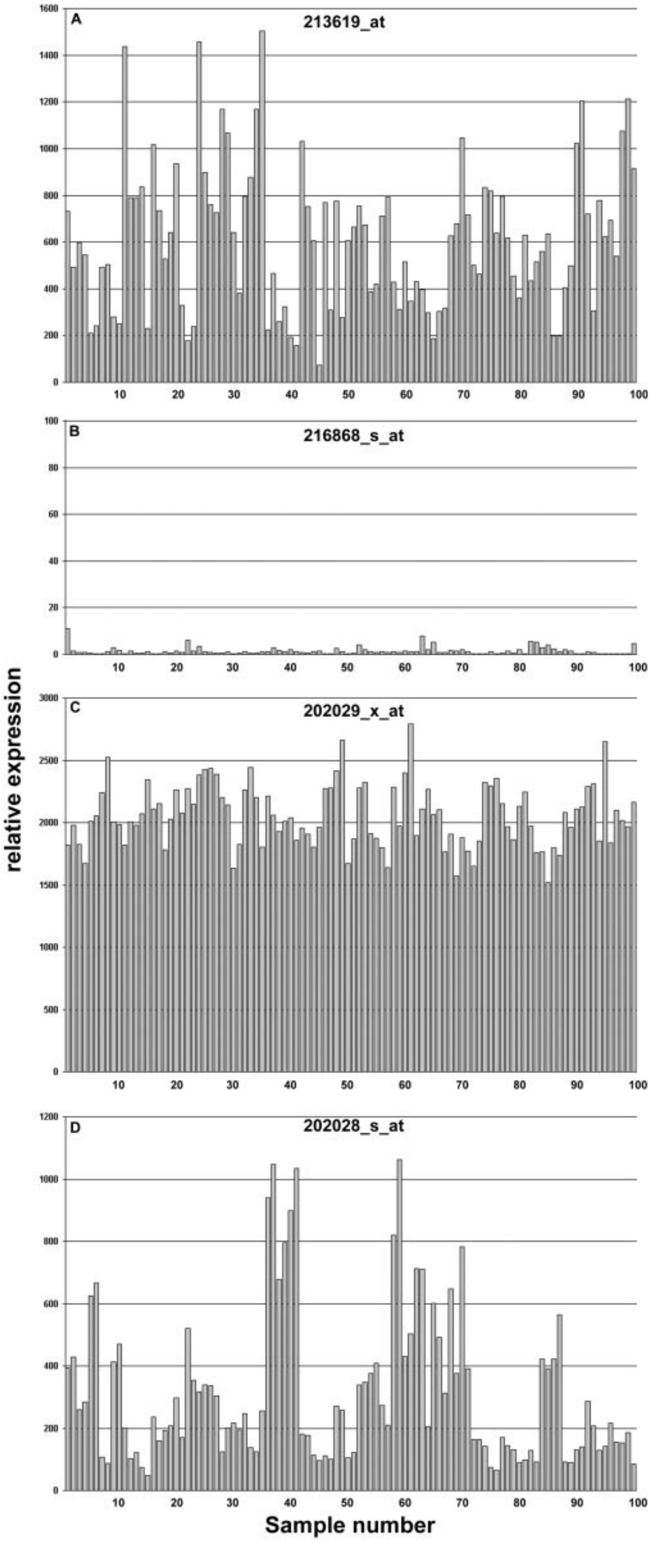

Figure 1. The signal captured by some probesets on the U133A array from 100 RNA samples collected from various tissues. Probeset 202029_x_at detects the expression of ribosomal protein L38. The other three probesets were designed to the complementary strand of the intended reference gene. Probeset 202028_s_at detects sequences complementary to the ribosomal protein L38. The plots for probesets 213619_at and 216868_s_at illustrate the difference between a probeset that detects a transcript and a probeset that does not detect a transcript. Although each plot is represented against a different scale, the relative expression levels are directly comparable.

the nature of this transcript is unknown. We have indicated that probesets such as this detect no known gene in our identification and therefore investigators will have to carefully consider what might be detected by these probesets. As our understanding of the human genome matures, these transcripts might become known, but for now the identification is uncertain. Using the expression level of samples from diverse cells and tissues can sometimes yield surprising results. Figure 1 also shows the expression level across 100 tissues of the ribosomal protein L38. This ribosomal protein is constitutively and consistently expressed. There are also two probesets on the U133 plus 2.0 array that detect the complementary strand to this gene. Probeset 202028_s_at is shown in Figure 1. Probeset 221943_x_at demonstrates the exact expression pattern (data not shown). The results indicate that in some cell types there is a transcript produced either from the opposite strand at the ribosomal protein L38 locus or from one of the many pseudogenes for RPL38 found in the human genome. This transcript, and its role in cells, has not yet been described. This kind of discovery is only possible with a careful analysis of what the probes detect.

Mixed probesets can complicate identification

We have identified 18 probesets that we define as chimeric probesets. Some of the probes in the probeset detect one gene and the rest detect a distinctly different gene. These appear to have been designed based on chimeric sequences like those described above. We have also found 153 probesets that contain at least one probe that hybridizes to ALU sequences. ALU sequences are highly repetitive elements that are found in the genomes of all apes (22). The sequence is a component of many genes and is incorporated into many transcripts. The probeset designed specifically to detect this sequence (AFFX-hum_alu_at) is always one of the highest intensity probesets on any array. Individual probes that cross-hybridize with ALU sequences can compromise the values calculated for gene expression because the signal can come from so many diverse transcripts. We have also found a number of probesets that contain one or two probes that do not detect any human transcripts. Since the analysis algorithms are designed to use all probes in a probeset, this can lead to a diminished performance of the probeset as a group.

Gene annotations should contain indications of uncertainty

The probeset annotation that we are advocating is represented in Table 4. This table illustrates some of the cautions that one must consider when identifying the 'gene' detected by a probeset on any array. Some probesets detect multiple genes or multiple splice variants from a single locus such as 201002_s_at and 203639_s_at respectively. We have identified other probesets on the array that can also identify these or related transcripts. It may be possible to narrow the field of possible transcripts by reviewing the behavior of all the related probesets as a group. Occasionally, a probeset detects more than one transcript and no mechanism exists to reduce the choices further, such as with probeset 202039_at. In this case, one will have to resort to some other experimental system to determine the correct gene changed in the microarray experiment performed. In some probesets, only a few of the probes may efficiently bind to multiple human transcripts. This is illustrated by probeset 206900_x_at. In this case, one might wish to look at the probe level data to determine if the changes are attributed to the primary gene or caused by cross-hybridization of other transcripts. Although it is not indicated in Table 4, we have the probes that might cross-hybridize in the full database. Some probesets do not detect a gene very effectively because of problems in the probe design. This is represented by probeset 217547_x_at in Table 4. In this case, because of inefficient binding of many of the probes to the mRNA target, the signal from the microarray should be viewed with caution. Probeset level data is based on the performance of 11 or more probe pairs and when only a few of the probes are hybridizing, it is unclear if the result is a consequence of the probes that hybridize to target or the probes that hybridize nonspecifically to other RNAs. Therefore, one must view the results with caution and possibly rely on other information. Some probesets do not match well with any known human transcripts. Most of these are the result of probes designed based on EST sequences. The U133 plus 2.0 array contains a large number of probesets that were designed based on EST sequences. There are nearly 11 000 probesets that do not match a described human gene. In many situations, the probes recognize sequences within the genetic locus of a known gene, but not sequences found in documented human transcripts. One example of this is shown for probeset 211610_at. Since we really know very little about the human transcriptome, it is not possible to completely disregard these probes. Many of these may be identified, as our knowledge progresses. Currently, we have indicated that the probeset should be checked carefully. It is possible that the signal measured is due to noise, and it is equally possible that it is due to an undescribed transcript. The reasons behind the behavior observed for these probesets in a microarray experiment require more evaluation and possibly follow-up experiments by the scientist.

DISCUSSION

The improper identification of probes on a microarray can influence many aspects of a microarray experiment. Analysis methods identify probes and these probes are converted into gene names by some annotation method. A failure at the level of annotation is often a cause for later questioning the analysis method. Of more concern is the fact that most microarray papers only mention the gene name. If the names are wrong, it creates difficulty for those who would later attempt to reproduce the results in alternative formats. This 'poor reproducibility' is the reason that validation is often required following microarray experiments. But validation will only be successful if one carefully evaluates what must be validated. The simplest case to consider is where several splice variants can be produced from a genetic locus and a spot on an array detects all variants. One must consider whether all transcripts, as a collective group, caused a detected change, or whether a single splice variant was responsible for the change. The former case might indicate a change in the initiation of transcription, while the latter case argues for splicing factors or transcript stability as the cause. Too often one assumes that the cause was transcriptional activation and that all methods

Table 4. Representative annotation of several Affymetrix probesets from the U133 plus 2.0 GeneChip

Probeset ID	Best matches (average score)	Multiple genes	Splice variants	Related probesets	Unigene number	Gene name	Gene symbol	Entrez Gene ID
201002_s_at	NM_021988 (25.0)	√	√	201003_x_at 208270_s_at	Hs.381025	Ubiquitin-conjugating enzyme E2 variant 1, transcript variant 1	UBE2V1	7335
	NM_199144 (25.0)			201001_s_at 210886_x_at	Hs.381025	Ubiquitin-conjugating enzyme E2 variant 1, transcript variant 2	UBE2V1	7335
	NM_022442 (25.0)			210241_s_at 216315_x_at	Hs.381025	Ubiquitin-conjugating enzyme E2 variant 1, transcript variant 3	UBE2V1	7335
	NM_003349 (25.0)				Hs.381025	Ubiquitin-conjugating enzyme E2 Kua-UEV isoform 2	Kua-UEV	387 522
	NM_199203 (25.0)				Hs.381025	Ubiquitin-conjugating enzyme E2 Kua-UEV isoform 1	Kua-UEV	387 522
202039_at	NM_004740 (25.0)	√	√		Hs.354085	TGFB1-induced anti-apoptotic factor 1	TIAF1	9220
	NM_078471 (25.0)				Hs.354085	Myosin XVIIIA	MYO18A	399 687
206900_x_at	NM_021047 (25.0)	√		221625_at hum_alu_at 206572_x_at 217547_x_at 215532_x_at 221626_at 215758_x_at	Hs.407162	Zinc finger protein 253	ZNF253	56 242
203639_s_at	NM_000141 (25.0)		√	AFFX-hum_alu_at	Hs.404081	Fibroblast growth factor receptor 2, transcript variant 1	FGFR2	2663
	NM_022969 (25.0)			211401_s_at 203638_s_at	Hs.404081	Fibroblast growth factor receptor 2, transcript variant 2	FGFR2	2663
	NM_022970 (25.0)			208225_at 208228_s_at	Hs.404081	Fibroblast growth factor receptor 2, transcript variant 3	FGFR2	2663
	NM_022972 (25.0)			208234_x_at	Hs.404081	Fibroblast growth factor receptor 2, transcript variant 5	FGFR2	2663
	NM_022975 (25.0)				Hs.404081	Fibroblast growth factor receptor 2, transcript variant 8	FGFR2	2663
	NM_023028 (25.0)				Hs.404081	Fibroblast growth factor receptor 2, transcript variant 10	FGFR2	2663
	NM_023029 (25.0)				Hs.404081	Fibroblast growth factor receptor 2, transcript variant 11	FGFR2	2663
	NM_023030 (25.0)				Hs.404081	Fibroblast growth factor receptor 2, transcript variant 12	FGFR2	2663
	NM_023031 (25.0)				Hs.404081	Fibroblast growth factor receptor 2, transcript variant 13	FGFR2	2663
1007_s_at	NM_001954 (25.0)		√	207169_x_at 210749_x_at	Hs.423573	Discoidin domain receptor family, member 1, transcript variant 2	DDR1	780
	NM_013993 (25.0)			208779_x_at	Hs.423573	Discoidin domain receptor family, member 1, transcript variant 1	DDR1	780
	NM_013994 (25.0)				Hs.423573	Discoidin domain receptor family, member 1, transcript variant 3	DDR1	780
217547_x_at	NM_007153 (5.83)	√			Hs.515712	Zinc finger protein 208	ZNF208	7757
211610_at	No best match				Hs.534315	Caution, check this probeset carefully. This probeset may detect an unusual splice variant, alternate termination site, or extended transcript of core promoter element binding protein. It is also a chimeric probeset with some of the probes detecting a locus 1.7 Mb away on chromosome 10	COPEB	1316

are equivalent for measuring this phenomenon. The proper identification of the targets for each probe and the careful examination of the performance of multiple-related probes can provide more information about the events occurring within the experimental system. This kind of care with gene identification can lead to more productive follow-up experimentation and greatly improve the verification of microarray results.

Improper identification comes from several sources. Not knowing the actual sequence placed on a spot makes proper identification impossible (10). Even when the sequence is known, identification can be uncertain, as many poorly identified sequences have been deposited in GenBank. Furthermore, the assumption that a probe detects only one sequence leads to errors. If multiple transcripts can bind to a spot but only one name is given, then sometimes the identification will be correct and sometimes it will be wrong. This problem might be overcome by evaluating the performance of several related probes to discern which gene has changed. But this is not done when one assumes that each spot corresponds to a single gene or that duplicate probes are equivalent. Finally, the complexity of the human genome is often ignored when deciding what a probe has detected. Splice variants and overlapping transcripts are usually ignored when naming spots and the user of the array often grabs the first sequence they find with the same name as that designated for the spot. Most of these problems are created by the desire to handle thousands of measurements quickly and efficiently. This is the ultimate source of error in a microarray experiment: compromising scientific accuracy for speed and ease of use.

We have re-identified the probes on the Affymetrix U133 plus 2.0 array. We have done this based on the sequence of the probes. In redefining the genes detected, we have removed two sources of error in the identification of probes on a microarray, and made it possible to do the proper analysis required to remove other sources of misidentification. The most important attribute of our identification of the U133 plus 2.0 probesets is that we indicate when the identification is uncertain. We know very little about the human genome and how it leads to a functional organism. Therefore, most of the identifications have a degree of uncertainty. However, we have only flagged genes for specific reasons. The possible reasons include the following: probesets that correspond to a gene where splice variants are known; probesets that recognize more than one gene transcribed from different loci; chimeric probesets where some of the probes detect one gene and some of the probes detect a different gene or something not found in the human genome; probesets where some or all of the probes detect a repetitive element within the genome; and probesets that detect a GenBank sequence that is not yet well characterized by the scientific community including ESTs and hypothetical sequences.

When a probeset is flagged as uncertain more care is required in defining the gene and planning follow-up experiments. We recommend a thorough *in silico* analysis of the gene or genes identified by such probesets. We also recommend that the investigator go back to the original data and evaluate the performance of all related probesets. If three spots all detect the same thing and only one spot has changed during the experiment, then this change is probably noise. More importantly, if the three similar probesets detect different splice variants and only one has changed then one has more information about how the change might have occurred, what protein might be produced, and how to best verify the phenomenon. In our identification, simple flags indicate the need for such analysis. We additionally report the related probesets that might help one to identify the affected gene. The example we showed of the interferon alpha gene family illustrates that many of the probesets should be considered as groups, not as individuals.

At this point much of this analysis must be done by hand. This is a daunting prospect when one is faced with a list containing hundreds of probesets. The bioinformatics community can help with this effort in a number of ways. Simple programs can be designed to retrieve data from related probes once a probe is identified by an analysis method. More involved programs might contain the necessary rules for evaluating a family of probesets and draw the appropriate conclusions from the experimental data. Alternatively, it is possible to redefine the probes that comprise a probeset so that better evaluations could be done for specific transcripts. Ultimately, many of these problems can be reduced by using this information to design better probes.

The results of our analysis indicate that there are additional problems that might influence the analysis of a probeset. Some probes do not detect anything and would be expected to yield unusually low hybridization signals. Others detect highly repetitive elements found in many transcripts and are expected to give unusually high hybridization signals at all times. Outlier signal on even one probe can influence the calculations used to define the metrics of ‘detection’, ‘signal’ and ‘change’. Therefore, the interpretation of any experimental outcome should be done with an understanding of this complicating factor. At present, all algorithms used to calculate information from a probeset assume that all probes within the set detect the same gene. This analysis indicates that that is not always the case. At present, the uninformative probes could be masked from the analysis. In the future, the problematic probes could be replaced with more informative probes.

Many people try to develop a big picture view of their microarray results. They do this by tapping into information that has been associated with gene names. Array results are often sorted according to the gene ontologies assigned to the protein products or assigned to cellular pathways based on the function of the protein. This practice should be re-evaluated. Given that so many of the probesets detect multiple transcripts, is it really correct to select one gene name and assume that that was the one affected by the change observed on an array? Furthermore, many of these ontologies do not accurately represent the subtle differences in the behavior of the proteins encoded by transcriptional variants. Even though it requires more work, we believe that the effort should be spent on accurately identifying the targets detected in a microarray experiment before attempting to get a big picture view of the data.

ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by H. Lee Moffitt Cancer Center and Research Institute.

REFERENCES

1. Chuaqui,R.F., Bonner,R.F., Best,C.J., Gillespie,J.W., Flaig,M.J., Hewitt,S.M., Phillips,J.L., Krizman,D.B., Tangrea,M.A., Ahram,M. *et al.* (2002) Post-analysis follow-up and validation of microarray experiments. *Nature Genet.*, **32**(Suppl.), 509–514.
2. Ohyama,H., Zhang,X., Kohno,Y., Alevizos,I., Posner,M., Wong,D.T. and Todd,R. (2000) Laser capture microdissection-generated target sample for high-density oligonucleotide array hybridization. *Biotechniques*, **29**, 530–536.
3. Yang,I.V., Chen,E., Hasseman,J.P., Liang,W., Frank,B.C., Wang,S., Sharov,V., Saeed,A.I., White,J., Li,J. *et al.* (2002) Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol.*, **3**, research0062.
4. Luzzi,V., Mahadevappa,M., Raja,R., Warrington,J.A. and Watson,M.A. (2003) Accurate and reproducible gene expression profiles from laser capture microdissection, transcript amplification, and high density oligonucleotide microarray analysis. *J. Mol. Diagn.*, **5**, 9–14.
5. Iscove,N.N., Barbara,M., Gu,M., Gibson,M., Modi,C. and Winegarden,N. (2002) Representation is faithfully preserved in global cDNA amplified exponentially from sub-picogram quantities of mRNA. *Nat. Biotechnol.*, **20**, 940–943.

6. Dobbin,K.K., Beer,D.G., Meyerson,M., Yeatman,T.J., Gerald,W.L., Jacobson,J.W., Conley,B., Buetow,K.H., Heiskanen,M., Simon,R.M. *et al.* (2005) Inter-laboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clin. Cancer Res.,* in press.
7. Yue,H., Eastman,P.S., Wang,B.B., Minor,J., Doctolero,M.H., Nuttall,R.L., Stack,R., Becker,J.W., Montgomery,J.R., Vainer,M. *et al.* (2001) An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.*, **29**, e41.
8. Tsibris,J.C., Segars,J., Enkemann,S., Coppola,D., Wilbanks,G.D., O'Brien,W.F. and Spellacy,W.N. (2003) New and old regulators of uterine leiomyoma growth from screening with DNA arrays. *Fertil. Steril.*, **80**, 279–281.
9. Kothapalli,R., Yoder,S.J., Mane,S. and Loughran,T.P.,Jr (2002) Microarray results: how accurate are they? *BMC Bioinformatics*, **3**, 22.
10. Halgren,R.G., Fielden,M.R., Fong,C.J. and Zacharewski,T.R. (2001) Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res.*, **29**, 582–588.
11. Handley,D., Serban,N., Peters,D., O'Doherty,R., Field,M., Wasserman,L., Spirtes,P., Scheines,R. and Glymour,C. (2004) Evidence of systematic expressed sequence tag IMAGE clone cross-hybridization on cDNA microarrays. *Genomics*, **83**, 1169–1175.
12. Holloway,A.J., van Laar,R.K., Tothill,R.W. and Bowtell,D.D. (2002) Options available—from start to finish—for obtaining data from DNA microarrays II. *Nature Genet.*, **32**(Suppl.), 481–489.
13. Carvalho,B., Ouwerkerk,E., Meijer,G.A. and Ylstra,B. (2004) High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J. Clin. Pathol.*, **57**, 644–646.
14. Kane,M.D., Jatkoe,T.A., Stumpf,C.R., Lu,J., Thomas,J.D. and Madore,S.J. (2000) Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.*, **28**, 4552–4557.
15. Lockhart,D.J., Dong,H., Byrne,M.C., Follettie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Wang,C., Kobayashi,M., Horton,H. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
16. Knight,J. (2001) When the chips are down. *Nature*, **410**, 860–861.
17. Djerbi,M., Darreh-Shori,T., Zhivotovsky,B. and Grandien,A. (2001) Characterization of the human FLICE-inhibitory protein locus and comparison of the anti-apoptotic activity of four different flip isoforms. *Scand. J. Immunol.*, **54**, 180–189.
18. Pruitt,K.D., Katz,K.S., Sicotte,H. and Maglott,D.R. (2000) Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.*, **16**, 44–47.
19. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2003) NCBI reference sequence project: update and current status. *Nucleic Acids Res.*, **31**, 34–37.
20. Valmeekam,V., Loh,Y.L. and San Francisco,M.J. (2001) Control of exuT activity for galacturonate transport by the negative regulator ExuR in *Erwinia chrysanthemi* EC16. *Mol. Plant Microbe Interact.*, **14**, 816–820.
21. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
22. Jasinska,A. and Krzyzosiak,W.J. (2004) Repetitive sequences that shape the human transcriptome. *FEBS Lett.*, **567**, 136–141.
23. Gren,E., Berzin,V., Jansone,I., Tsimanis,A., Vishnevsky,Y. and Apsalons,U. (1984) Novel human leukocyte interferon subtype and structural comparison of alpha interferon genes. *J. Interferon Res.*, **4**, 609–617.
24. Tasic,B., Nabholz,C.E., Baldwin,K.K., Kim,Y., Rueckert,E.H., Ribich,S.A., Cramer,P., Wu,Q., Axel,R. and Maniatis,T. (2002) Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. *Mol. Cell*, **10**, 21–33.
25. Zhang,T., Haws,P. and Wu,Q. (2004) Multiple variable first exons: a mechanism for cell- and tissue-specific gene regulation. *Genome Res.*, **14**, 79–89.
26. Yoshihama,M., Uechi,T., Asakawa,S., Kawasaki,K., Kato,S., Higa,S., Maeda,N., Minoshima,S., Tanaka,T., Shimizu,N. *et al.* (2002) The human ribosomal protein genes: sequencing and comparative analysis of 73 genes. *Genome Res.*, **12**, 379–390.